The Revolution Will Not Be Supervised!

Self-Supervised Learning

Motivation

Approaches

Self-supervised Learning - Motivation

Motivation - the state of the (machine perception) nation

Reasons to be cheerful

Deep learning has achieved remarkable progress with supervised learning:

- Gather a large collection of data and manually annotate it
- Supervise a model with the resulting (data, annotation) pairs.

Major gains on vision benchmarks!

Causes for concern

Despite these successes, we still seem to have a long way to go:

- Even the *highest capacity* models trained on the *largest annotated datasets* continue to make "*silly*" mistakes
- It seems we can never get enough labelled data to get close to the human perception system

Can we take inspiration from the early stages of development of human perception?

Self-supervised Learning - Motivation

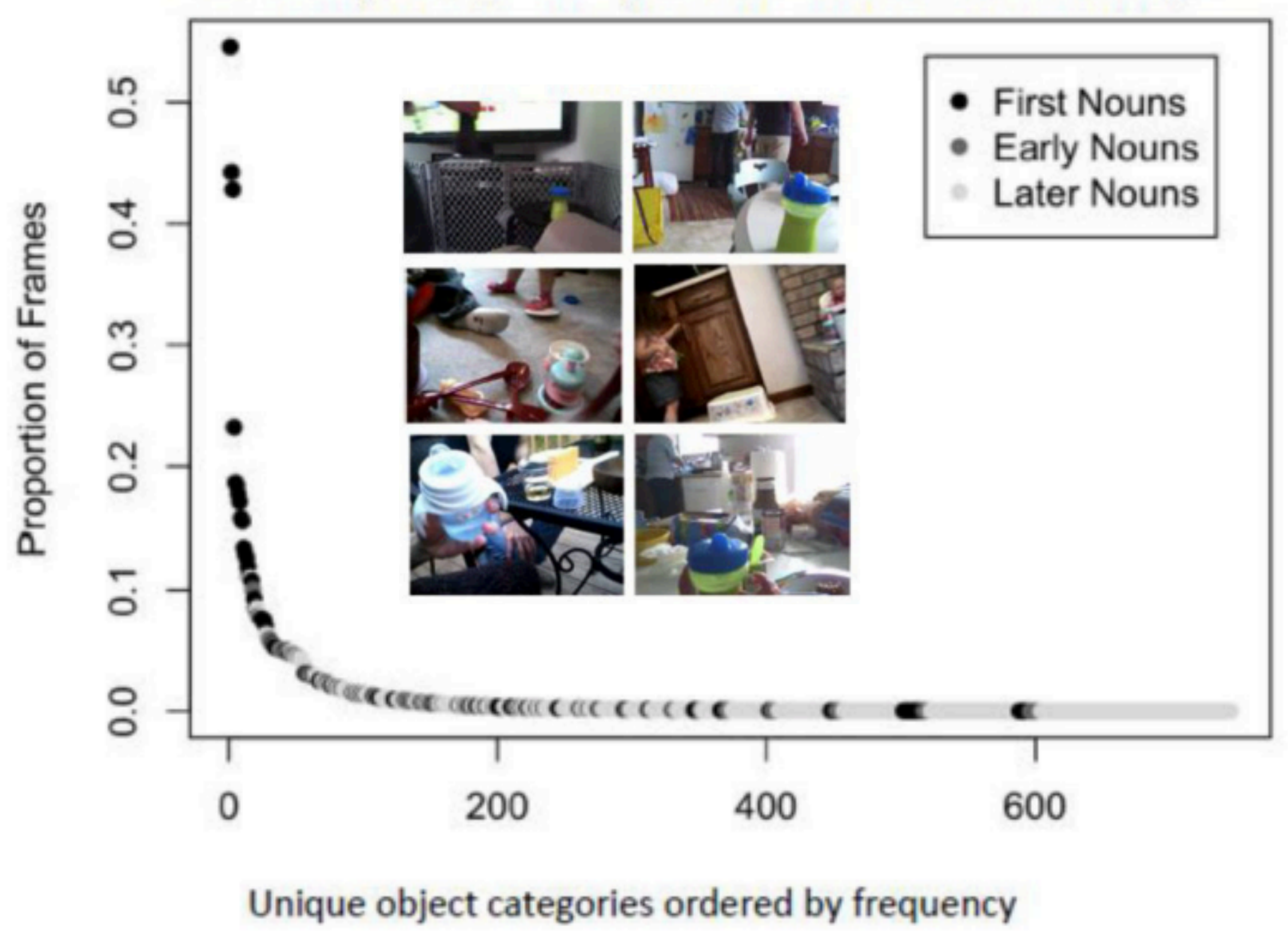
Lessons from Embodied Cognition

Human baby learning is:

- Incremental
- Social
- Physical
- Exploratory
- Language-based
- Multi-modal

We will discuss self-supervised methods (partly) inspired by human multi-modal learning (exploiting redundant signal)

Babies build curricula



Heavy focus on a small number of objects

Practical Challenges

"In order that the machine should have a chance of finding things out for itself it should be allowed to roam the countryside, and the danger to the ordinary citizen would be serious."

Turing, 1948

There are practical challenges to embodied learning

Simulation may help

References/Image credits

L. B. Smith and M. Gasser, "The Development of Embodied Cognition: Six Lessons from Babies," Artificial Life (2005)

L. B. Smith et al., "The Developing Infant Creates a Curriculum for Statistical Learning", *Trends in Cognitive Sciences* (2018)

A. M. Turing, "Intelligent Machinery", (1948)

Self-supervised Learning - creating your own supervision

Learning via prediction - Helmholtz

Each movement we make by which we alter the appearance of objects should be thought of as an experiment designed to test whether we have understood correctly the invariant relations of the phenomena before us, that is, their existence in definite spatial relations

Helmholtz, 1878

Generate labels by predicting the future

Redundancy provides knowledge - Barlow

To detect a new association (e.g. event C precedes event U), requires knowledge of the prior probabilities of C and U

We can then learn new associations as occurrences of C followed by U more frequently than would happen by chance

To know "what usually happens" we need **redundancy** in the input signal (e.g. views of the same event from different modalities)

Redundant signal (by definition) can be predicted from remaining signal

Generate labels from redundant signal

References/Image credits:
H. L. F. Helmholtz, "The Facts in Perception" (1878)
H. B. Barlow, "Unsupervised learning", Neural computation (1989)

Self-supervised Learning - creating your own supervision

Computational trick: factorial codes for learning new associations

When learning pairwise associations between N events, we need to store N^2 co-occurrence probabilities

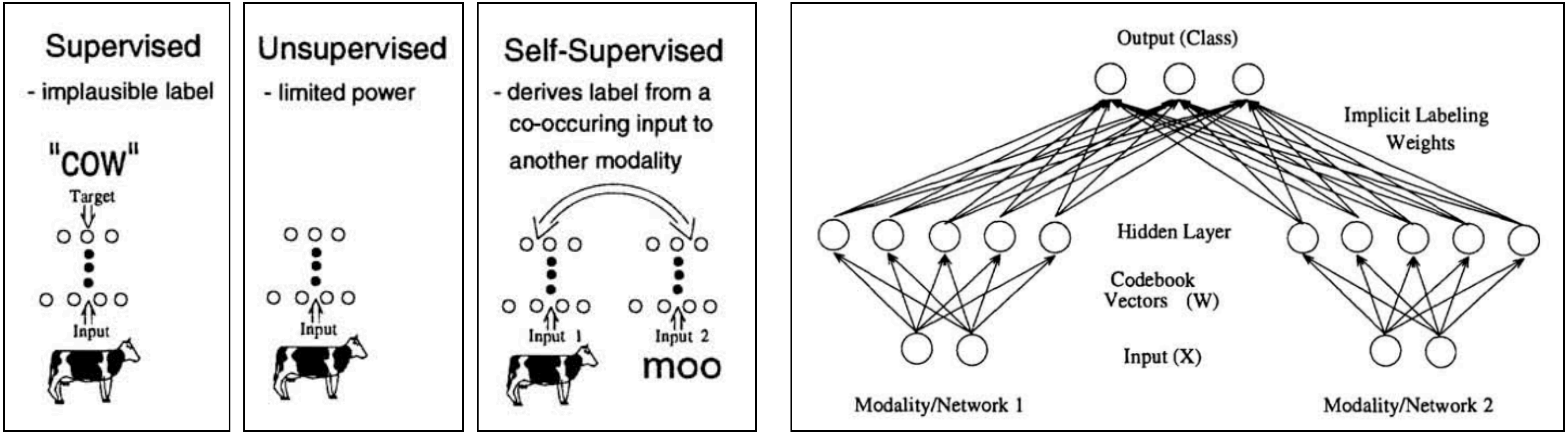
If our representations of events C and U are statistically independent, we can compute the chance co-occurrence of C and U from their marginals: i.e. $P(C)P(U)$, so we need only store N event probabilities!

Barlow suggested *Minimum Entropy Coding* to obtain such factorial representations - but this principle applies more generally

Reference:
H. B. Barlow, "Unsupervised learning", Neural computation (1989)

Self-supervised Learning - creating your own supervision

Exploiting Multi-modal Correlation - de Sa



Learning signal: Minimise disagreement between class labels predicted from each modality

Note: in modern research, distinction between self-supervised & unsupervised can be blurry....

References/Image credits:
V. R. de Sa, "Learning Classification with Unlabeled Data", *NeurIPS* (1993)

Self-supervised Learning - context as supervision

Natural Language Processing

Unlabelled text corpora have long been used to provide supervision for neural networks, with the hope that their distributed representations will enable generalisation

Autoregressive models

Factor the probability of a sequence, x_1^T, as conditionals:

$$P(x_1^T) = \prod_{t=1}^{T} P(x_t | x_1^{t-1})$$

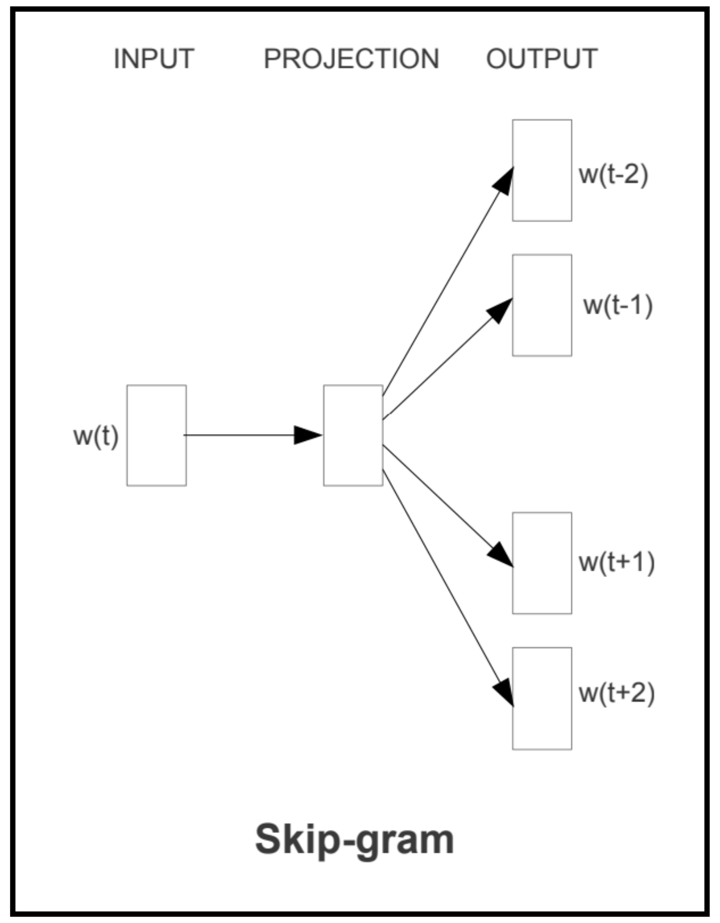
Maximise likelihood of text corpus

Predict next character (Schmidhuber et al., 1996)

Predict next word (Bengio et al., 2003)

Predicting context

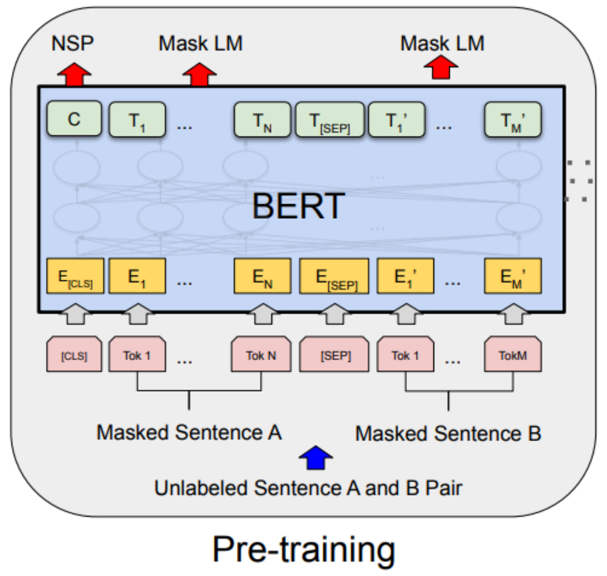
Word2Vec was trained to predict surrounding words



highlighted importance of having *lots* of training data

Multitask masking

BERT - trained to predict *randomly masked words* and next sentence prediction



Showed benefits of high-capacity bi-directional transformer

References/Image credits

J. Schmidhuber and S. Heil, "Sequential neural text compression", IEEE Trans. on Neural Networks (1996)

Y. Bengio et al., "A Neural Probabilistic Language Model", JMLR (2000)

T. Mikolov et al. "Efficient Estimation of Word Representations in Vector Space", ICLR (2013)

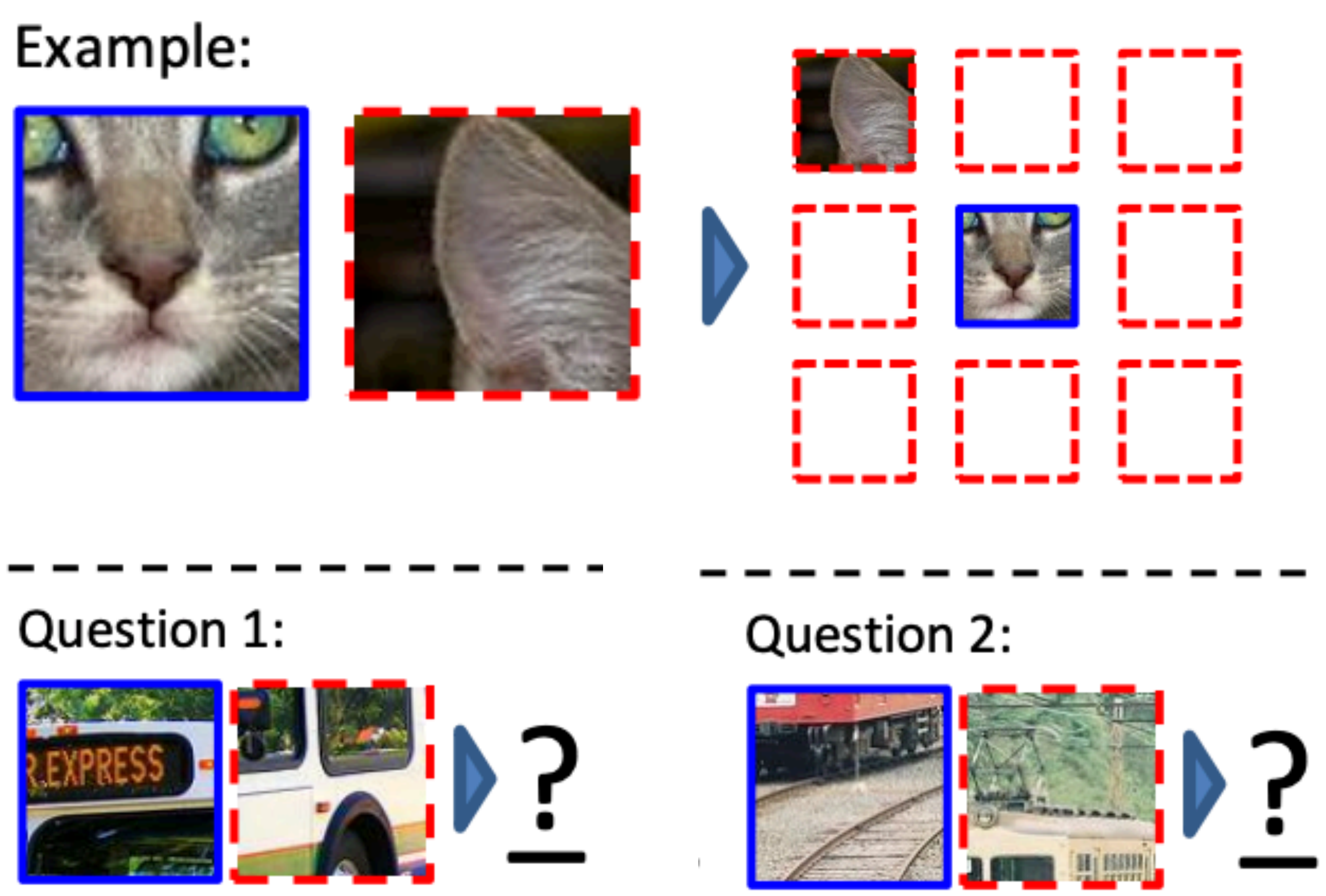
J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL (2019)

Back to Vision: context as supervision

Computer Vision

In vision, we train network by playing a game (often called a pretext task)

We typically don't care about performance on the pretext task itself, but we hope that by solving it, a model learns good representations of the visual world



Key idea: a model can only solve these questions once it learns about cats, buses and trains. No labelling is required!

Warning: sometimes the model won't solve the task in the way you wanted!

Doersch et al. found that the network could "cheat" by exploiting chromatic aberration to solve the puzzle unless it was prevented from doing so.

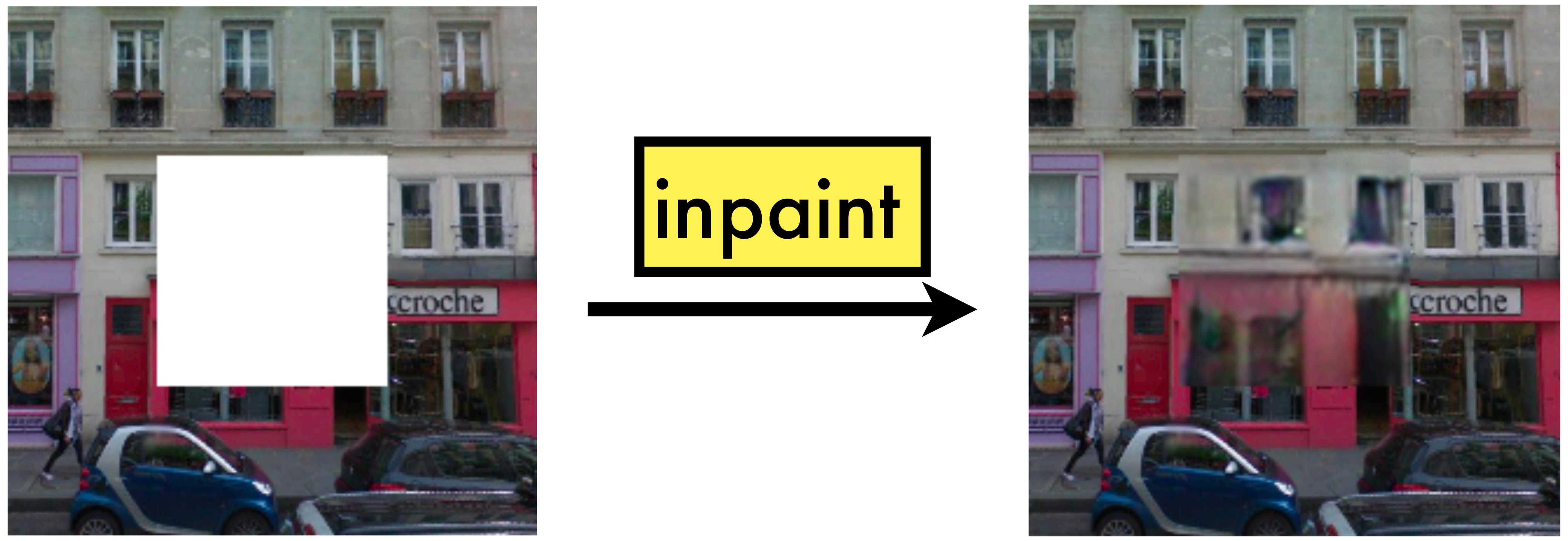
Note: also a problem for AI safety

References/Image credits

C. Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV (2015)

Pretext task: inpainting

Learning by Inpainting (Pathak et al., 2016)



Train model to "inpaint" (fill in the gap)

Loss contains two terms:

- $L2$ loss on patch reconstruction ($\mathbf{p}$ is a patch):

$$L_{\text{rec}} = ||\mathbf{p}_{\text{pred}} - \mathbf{p}_{\text{gt}}||_2^2$$

issue: blurry predictions

- Adversarial loss (inspired by GANs)

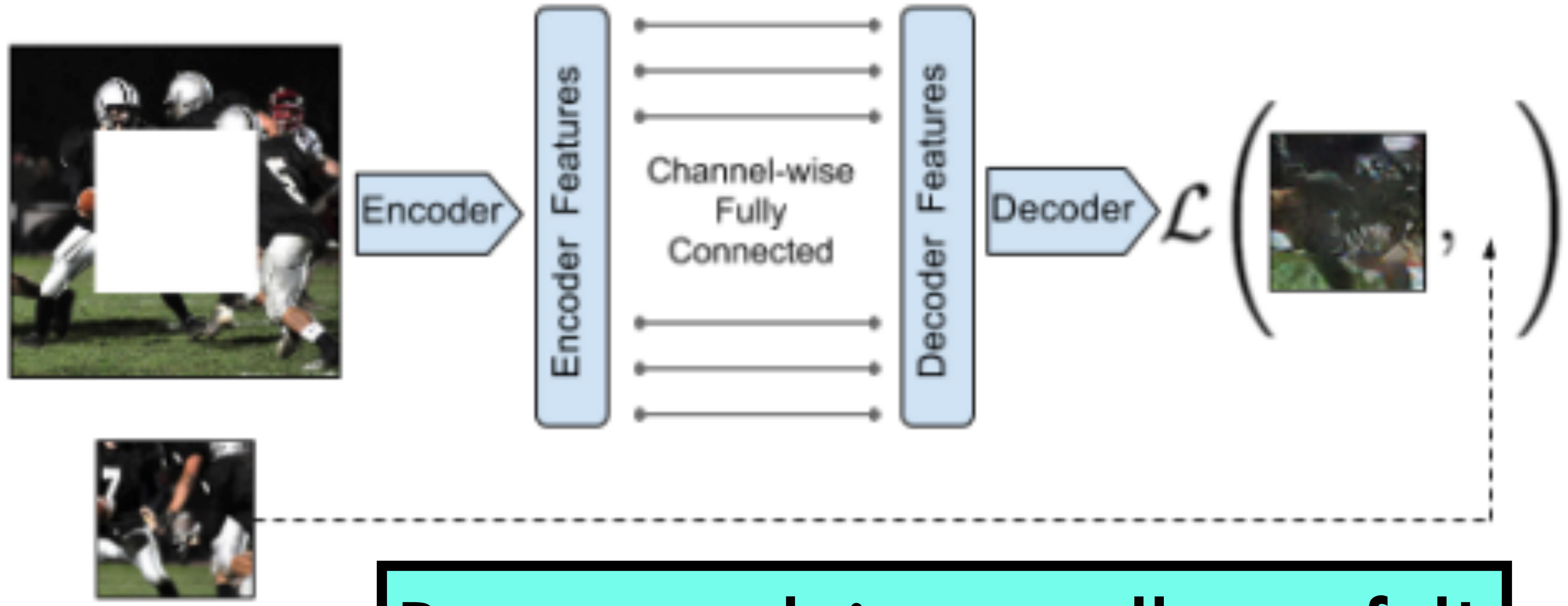
$$L_{\text{adv}} = \max_D \mathbb{E}_{\mathbf{p}_{gt}}[\log(D(\mathbf{p}_{gt})) + \log(1 - D(\mathbf{p}_{\text{pred}}))]$$

Train D to recognise real data

Train model to fool D

D is a second network ("the discriminator") trained jointly with main model

reduces blurriness



Pretext task is actually useful!

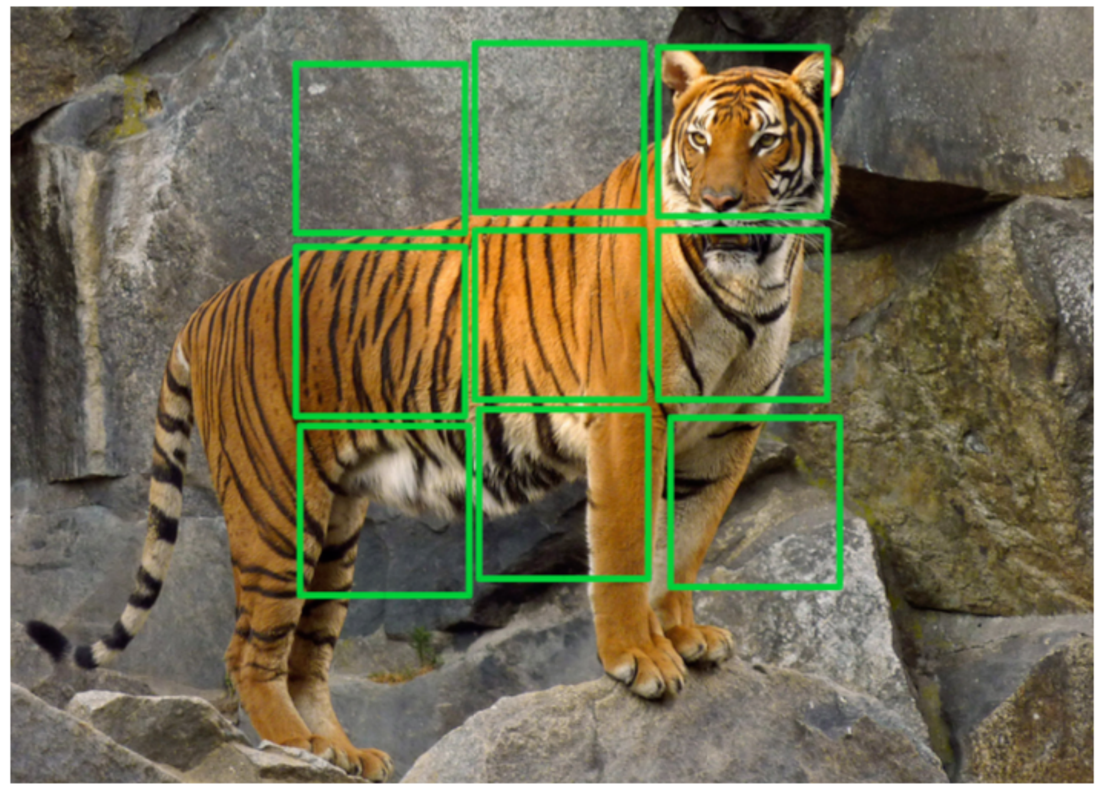
Encoder (w. L2 loss) learns useful features for classification, detection & segmentation

References/Image credits

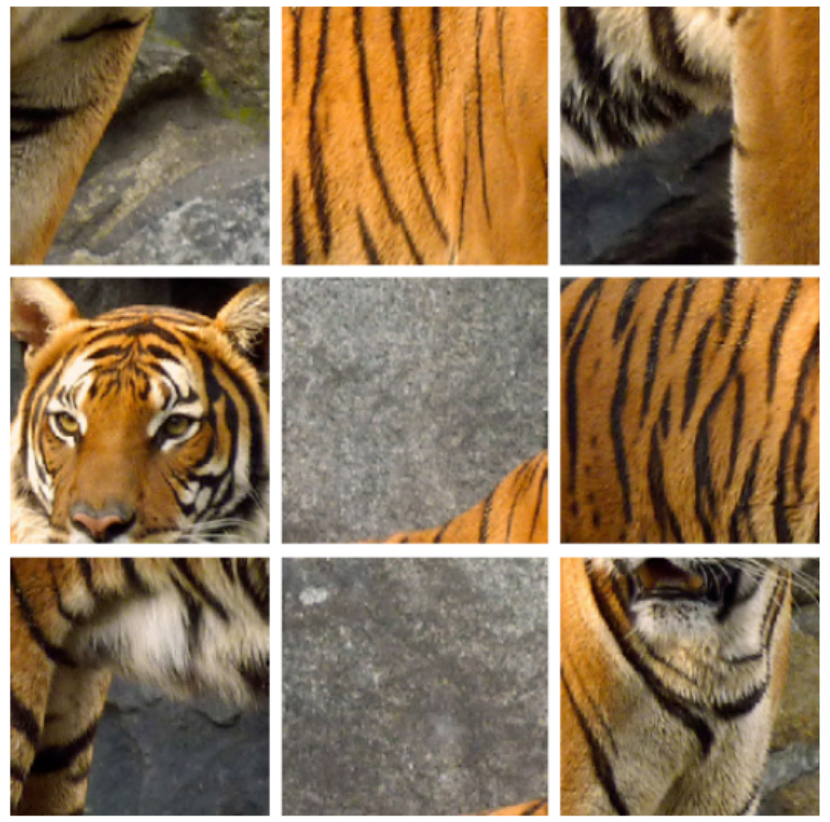
D. Pathak et al., "Context Encoders: Feature Learning by Inpainting", CVPR (2016)

Pretext task: jigsaw puzzles

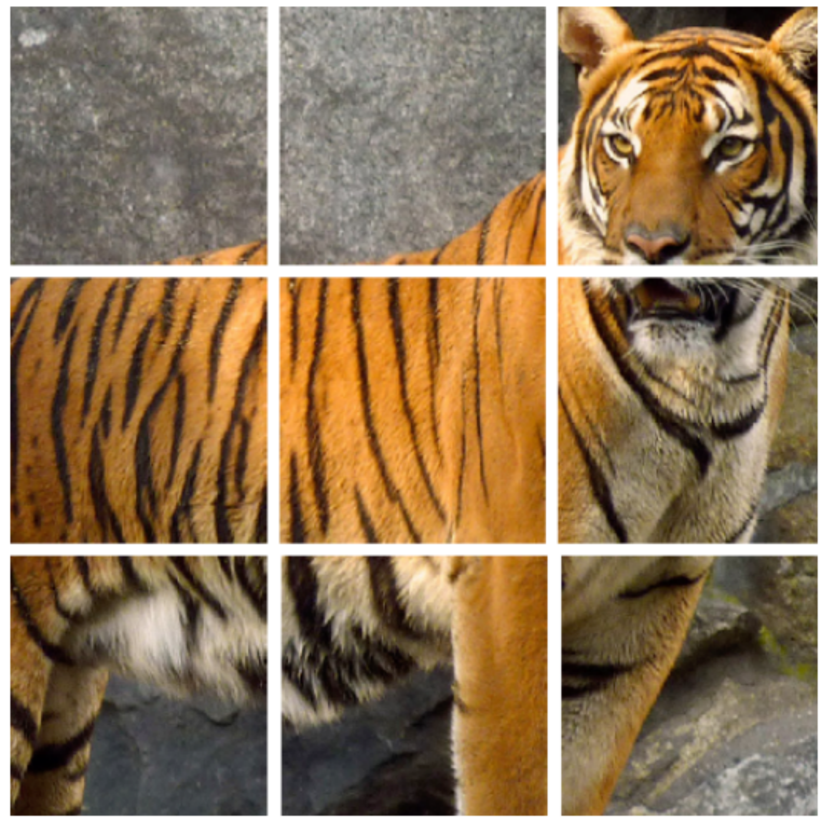
Learning from jigsaws (Noroozi and Favaro, 2016)

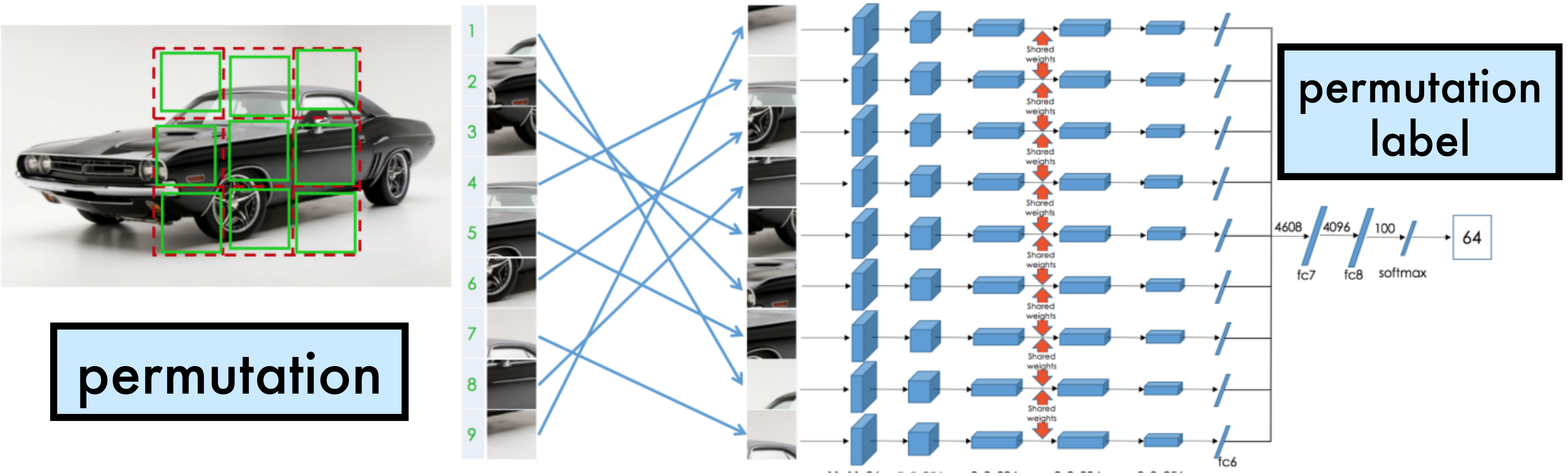


shuffle



solve



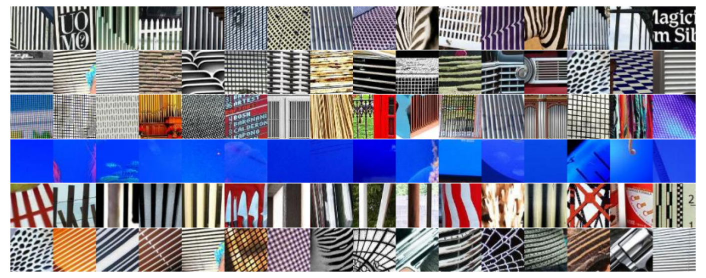


Features can be used for:

classification

image retrieval

Visualised activations:



Conv 1

Conv 4

References/Image credits

M. Noroozi and P. Favaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles", ECCV (2016)

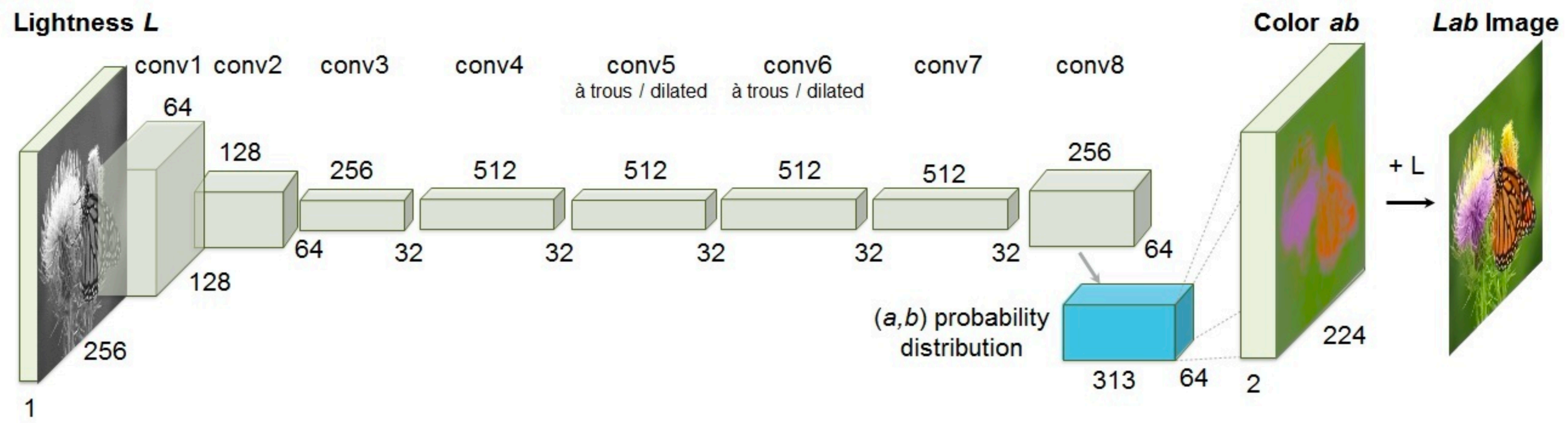
Pretext task: Colourisation

Learning from colourisation (Zhang et al., 2016)

colourise

Key premise:

The model can only fill in "plausible" colours if it understand the image



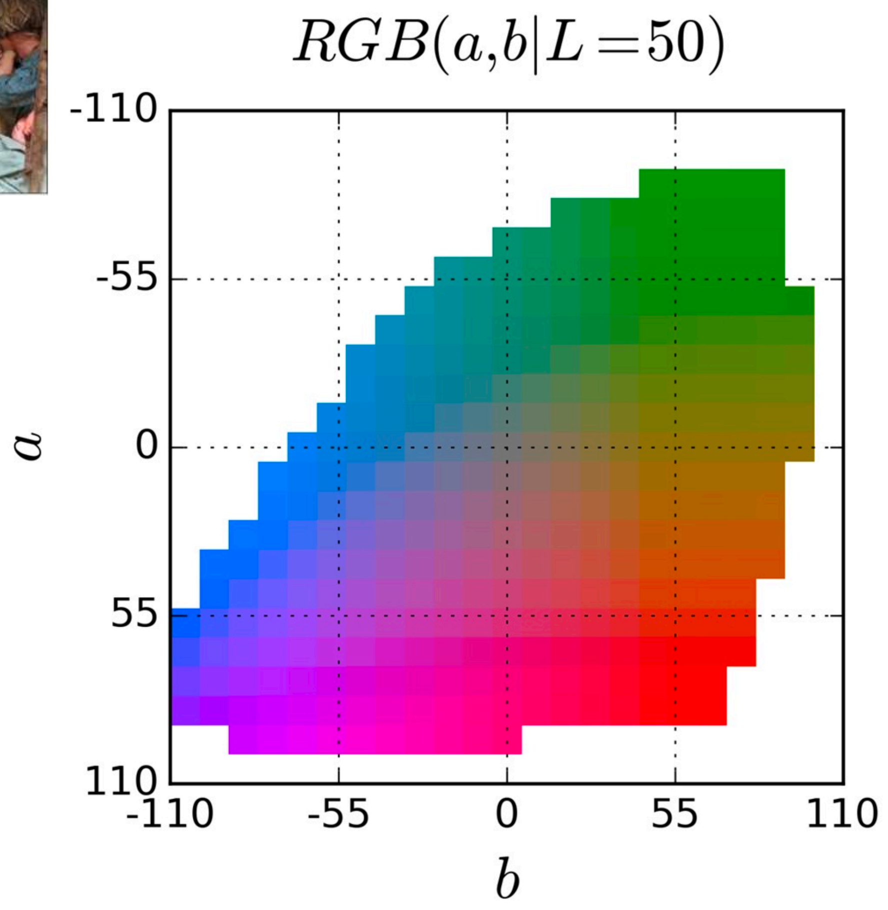
Challenge:

Colour distribution is multimodal 

L2 regression gives grey-ish colours

Solution:

Predict quantised **Lab** space values with cross entropy loss



Features can be used for:

classification | detection | segmentation

References/Image credits

R. Zhang et al., "Colorful Image Colorization", ECCV (2016)
https://phys.org/news/2017-07-images-deep-neural-networks.html

What's wrong with L2?

Learning in the presence of multimodal data

For many machine learning **unimodal** modelling problems, L2 regression is a good choice

However, it's not a good fit for predicting **multimodal** distributions

Why?

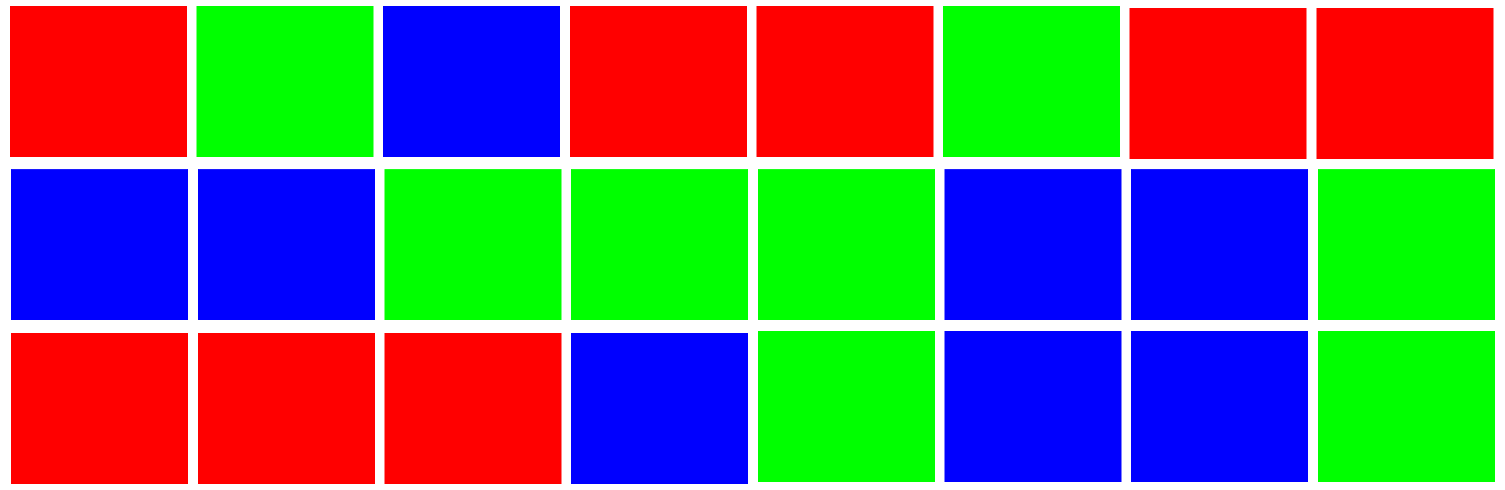
Suppose we wish to model a single RGB pixel image that:

Is pure red with probability $\frac{1}{3}$: RGB: (**1**, 0, 0)

Is pure green with probability $\frac{1}{3}$: RGB: (0, **1**, 0)

Is pure blue with probability $\frac{1}{3}$: RGB: (0, 0, **1**)

Training set



Loss: $||\mathbf{p}_{\mathrm{gt}} - \mathbf{p}_{\mathrm{pred}}||_2^2$

Optimal solution: (0.33, 0.33, 0.33)

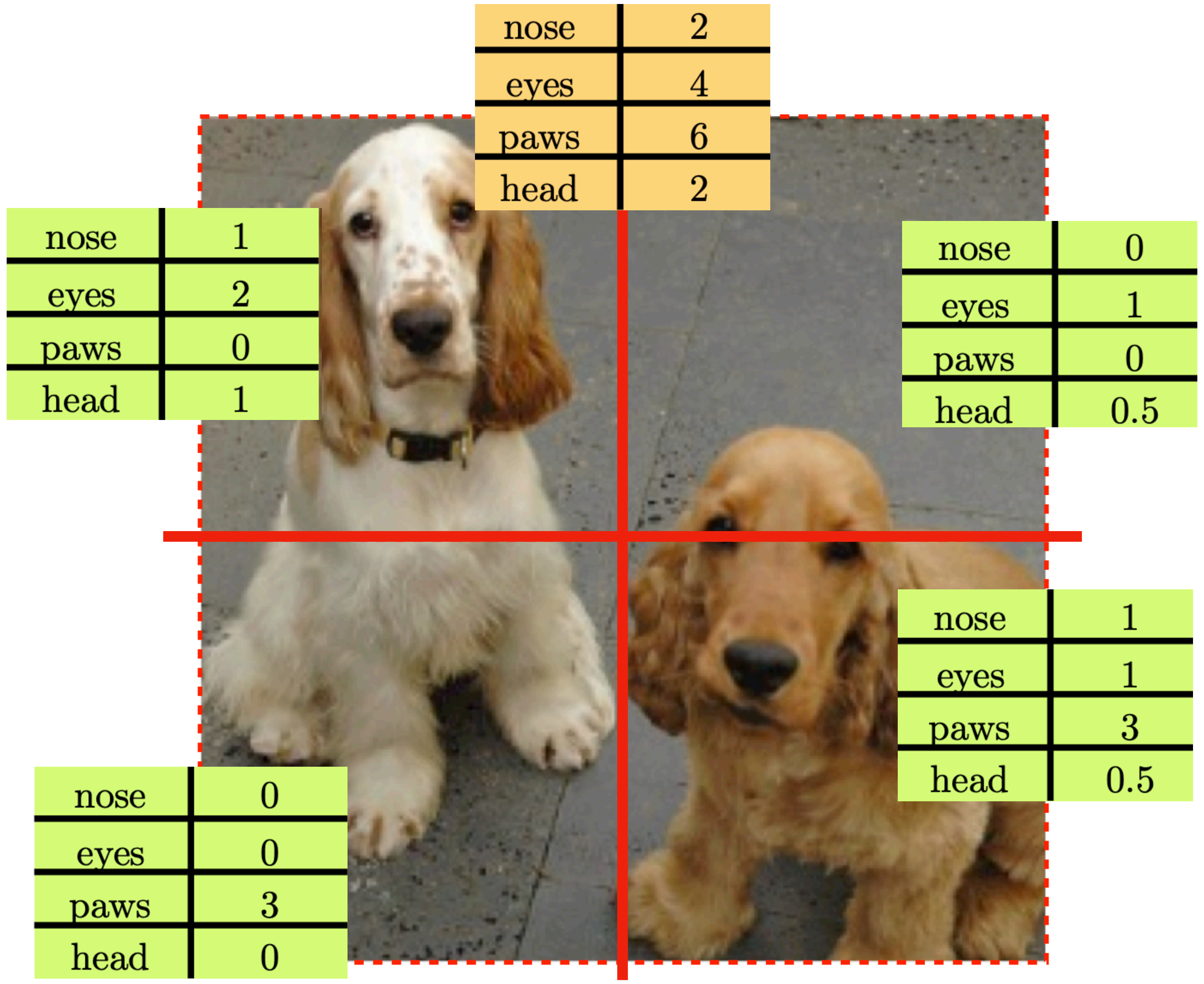
Samples from model 

References

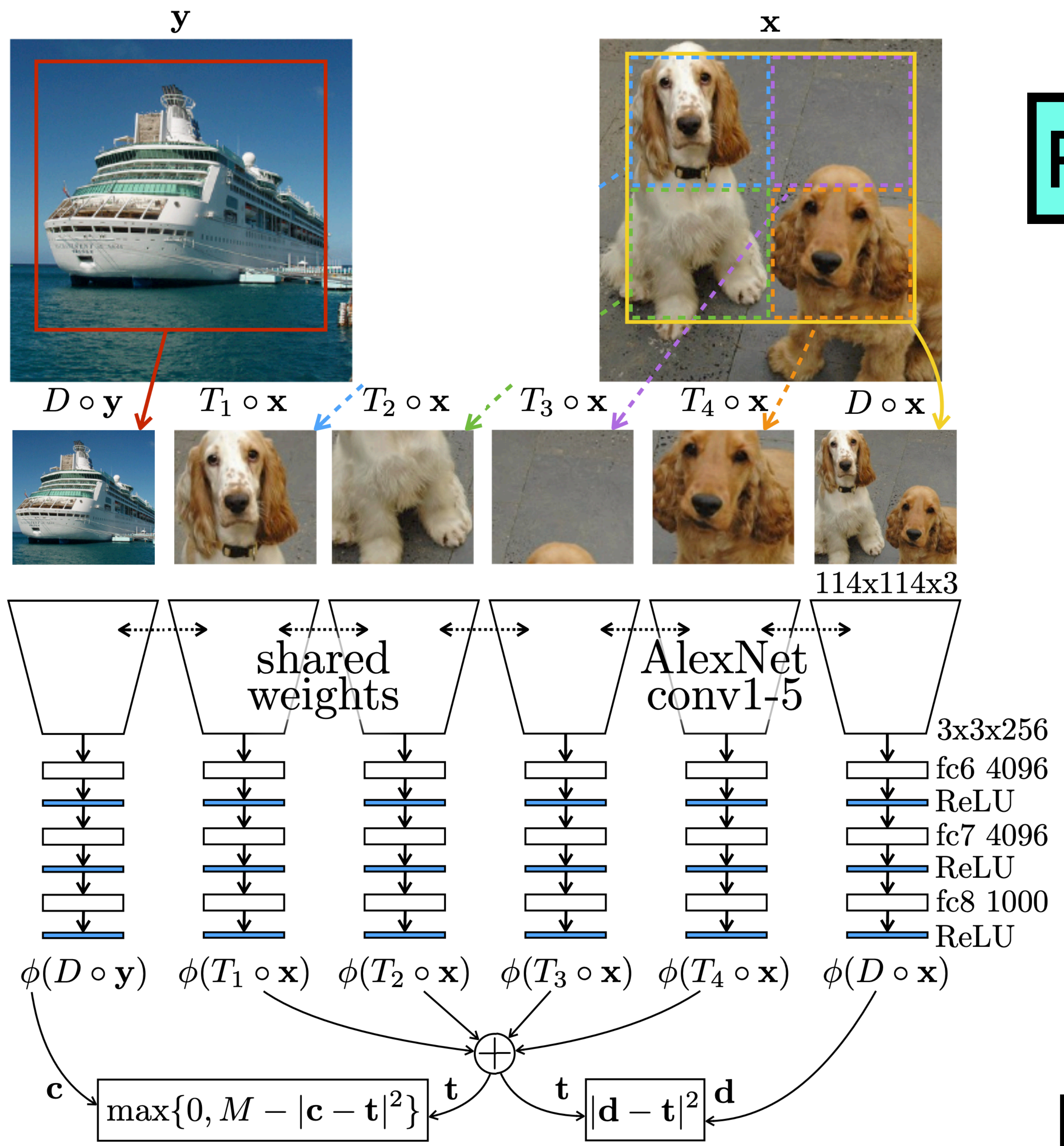
M. Mathieu et al., "Deep multi-scale video prediction beyond mean square error", arXiv preprint arXiv:1511.05440 (2015)

Pretext task: Counting

Learning from counting (Noroozi et al., 2017)



Consistency constraint



Problem: "Trivial solution"

Model can predict a count of 0 for every image we give it

Solution: add contrastive images and enforce different counts

Features can be used for:

classification | detection | segmentation

The counted concepts are not specified directly

(they are chosen by the model)

References/Image credits

M. Noroozi et al., "Representation Learning by Learning to Count", ICCV (2017)

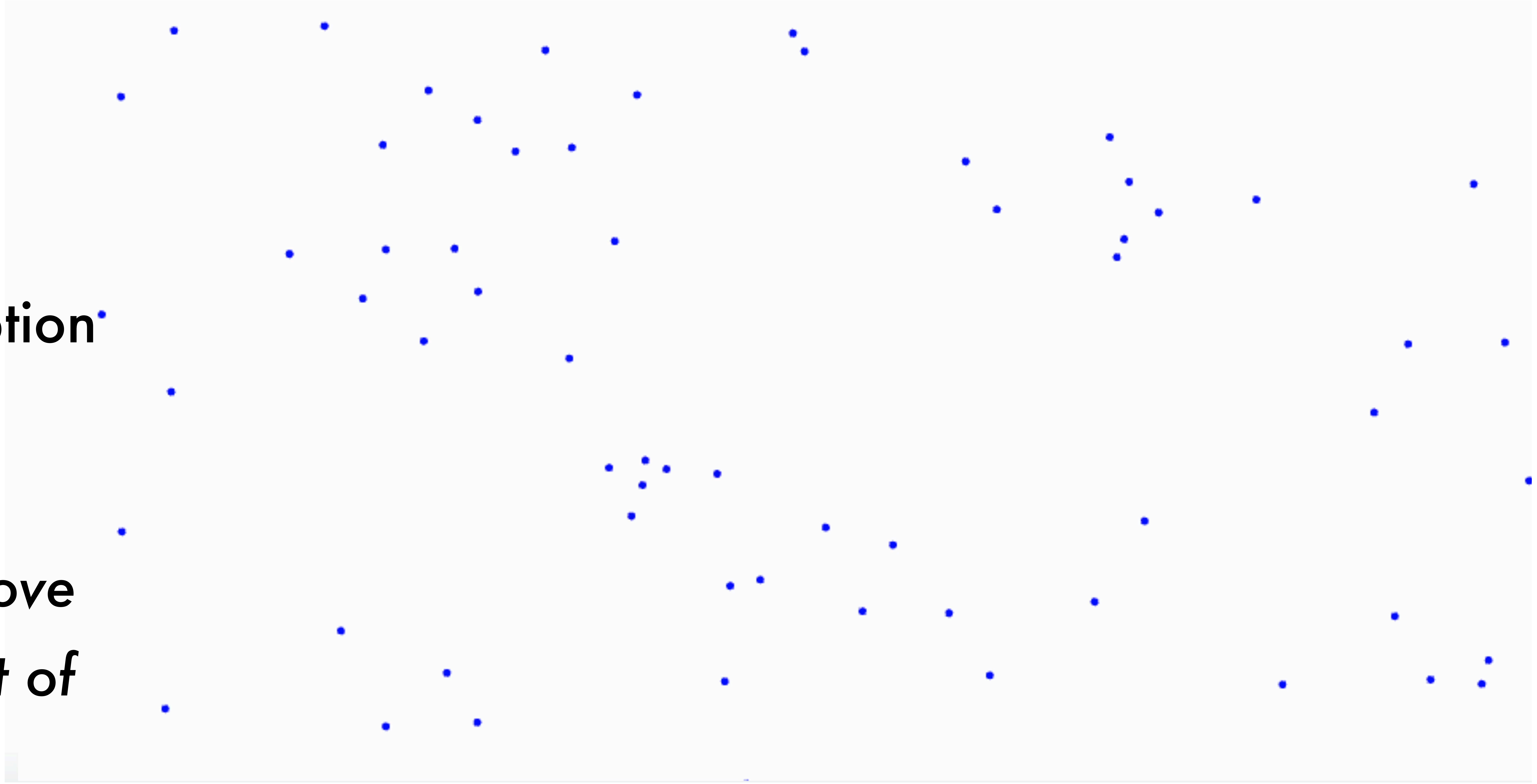
Grouping/Common Fate

Gestalt Principle: Common Fate

The Gestalt school of psychology emerged in the early 20th Century

It proposed several of "grouping principles" to explain human perception

The principle of "**common fate**":
We perceive visual elements that move with the same velocity as being part of a single whole



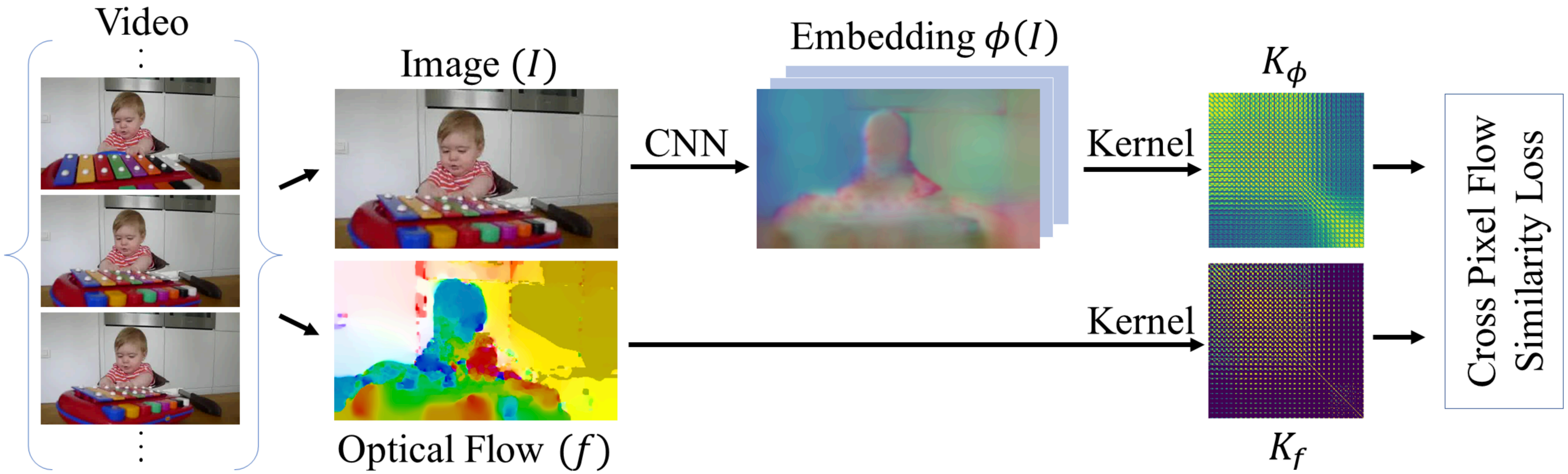
References/Image credits
Simulation: https://isle.hanover.edu/Ch05Object/Ch05CommonFate_evt.html

Pretext task: Grouping/Common Fate

Learning from Gestalt principles (Mahendran et al., 2018)

Key idea: pixels that belong to the same object are much more likely to "move together" than pixels that do not

Consistency constraint



Encourages network to learn embeddings of pixels that move together

Result: it learns to segment objects

Note: optical flow is a 2D vector field where each vector is a displacement vector showing the movement of points from one frame to another

Features can be used for:

classification | detection | segmentation

References/Image credits

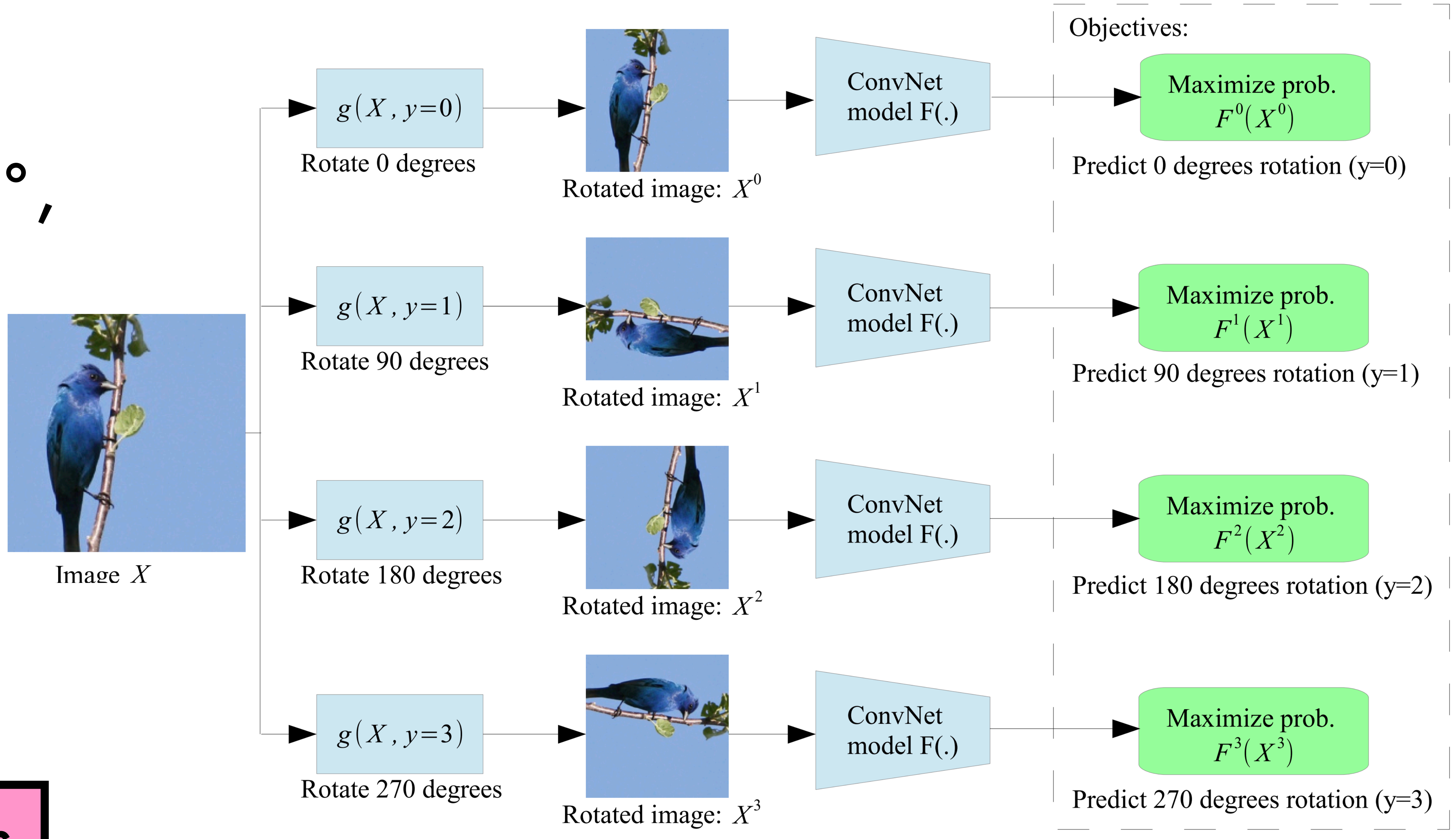
A. Mahendran et al., "Cross Pixel Optical Flow Similarity for Self-Supervised Learning", ACCV (2018)

Pretext task: Rotations

Learning from photographer bias (Gidaris et al., 2018)

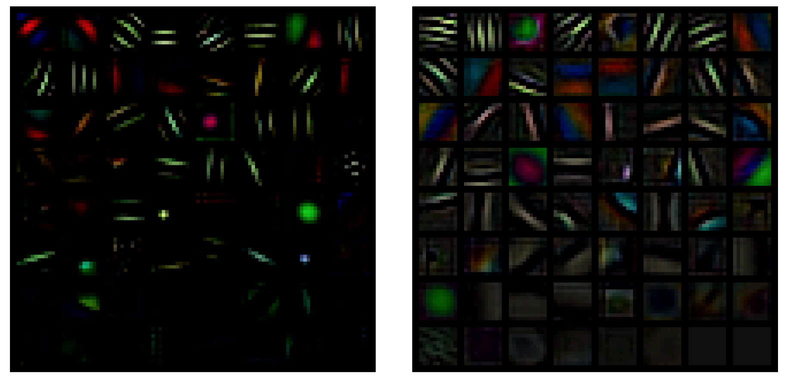
Humans take photos "the right way up"

If images are rotated anticlockwise by 0°, 90°, 180° or 270° we can spot the rotation



Empirically, this learns very strong features

First layer filters

Supervised  Self-supervised

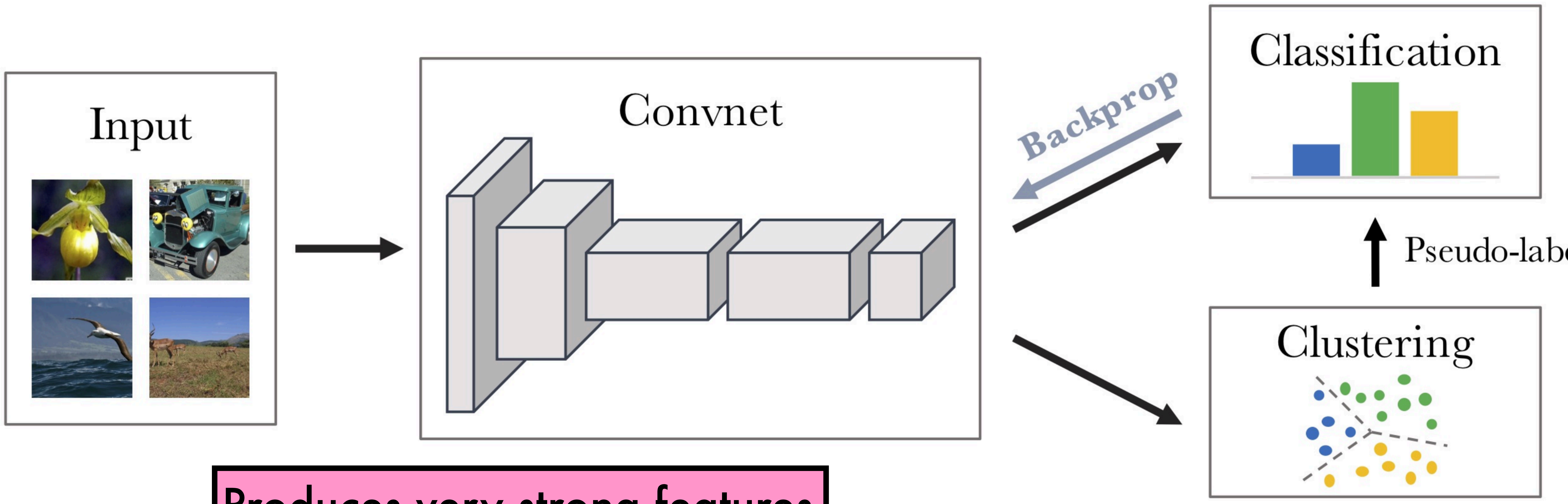
How? By understanding the image content

References/Image credits

S. Gidaris et al. "Unsupervised Representation Learning by Predicting Image Rotations", ICLR (2018)

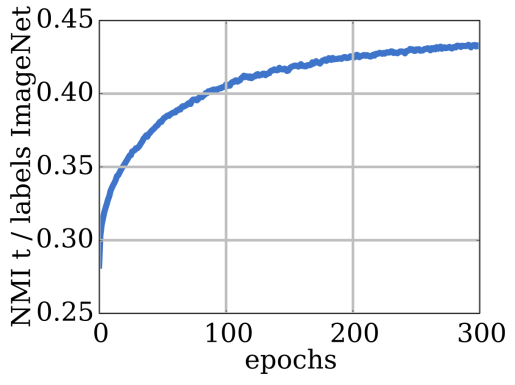
Pretext task: Clustering

Learning from clustering (Caron et al., 2018)

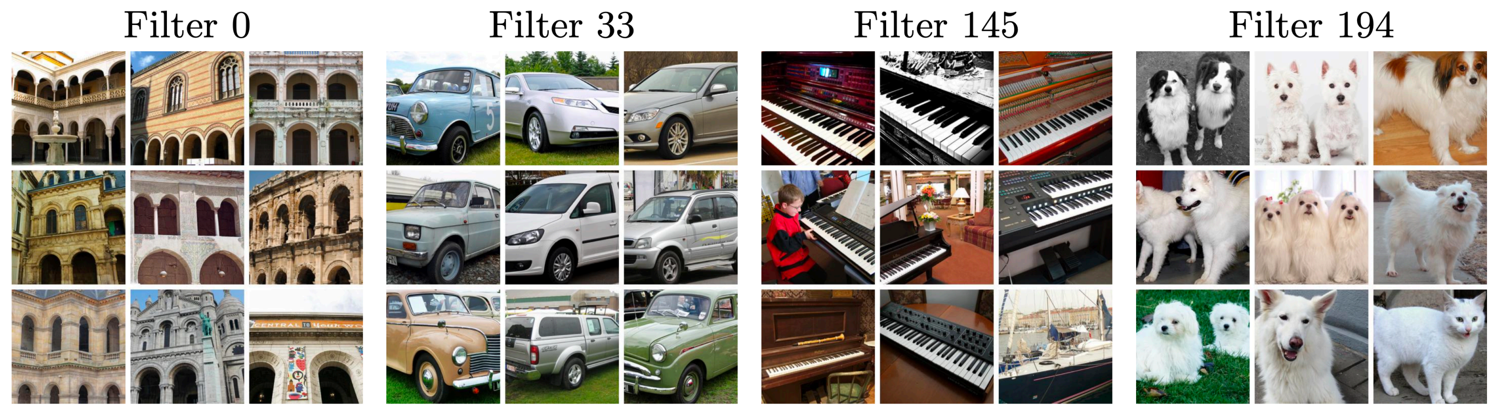


Produces very strong features

Note: Even with random weights, a CNN performs much better than random clustering



(a) Clustering quality



Visualised activations (last conv)

References/Image credits

M. Caron et al., "Deep Clustering for Unsupervised Learning of Visual Features", *ECCV* (2018)

Contrastive Learning

Learning from augmentations (Chen et al., 2020)

SimCLR: "Simple Framework for contrastive learning of visual representations"

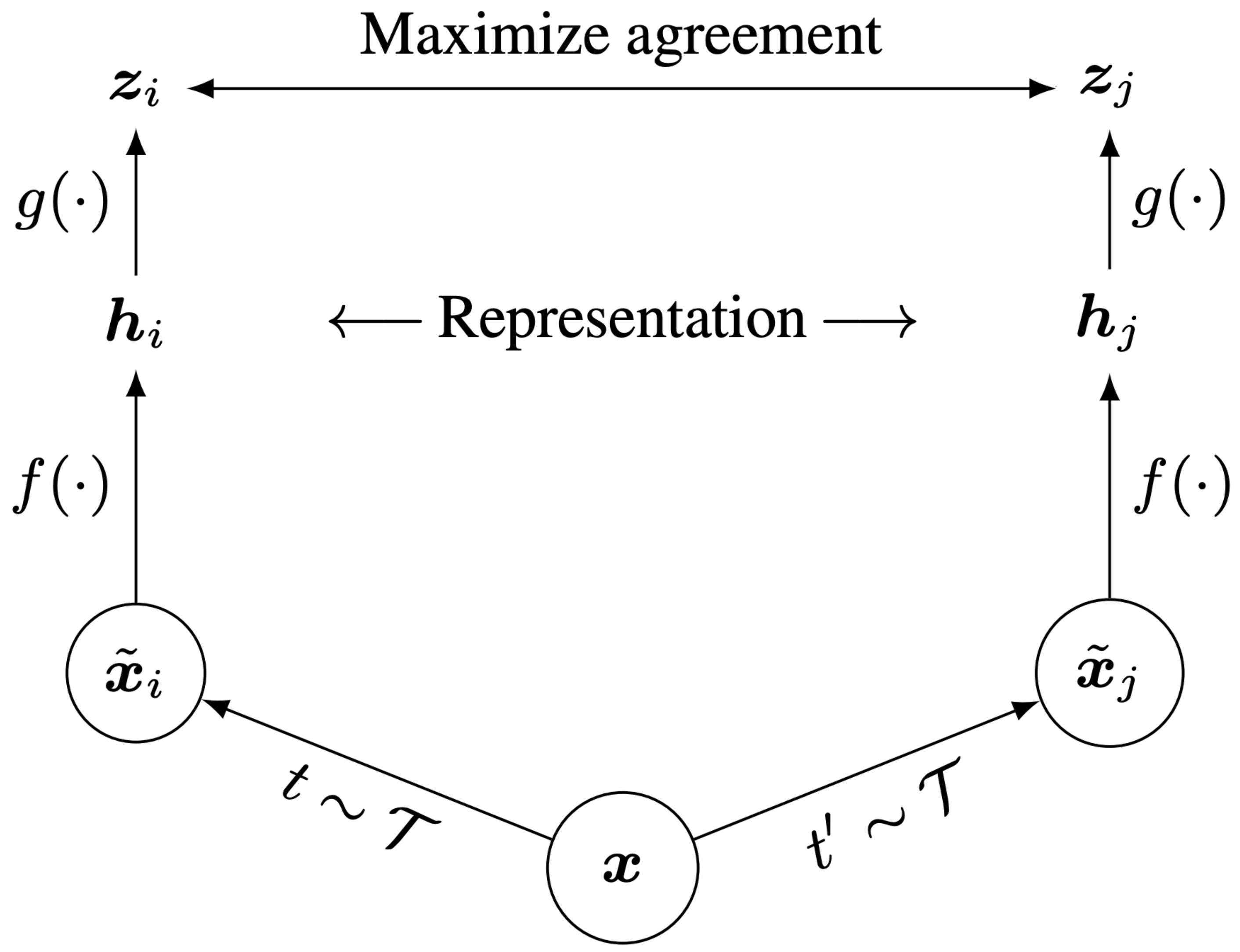
Idea: Data augmentation preserves semantic meaning


(a) Original


(b) Crop and resize


(c) Crop, resize (and flip)


(i) Gaussian blur



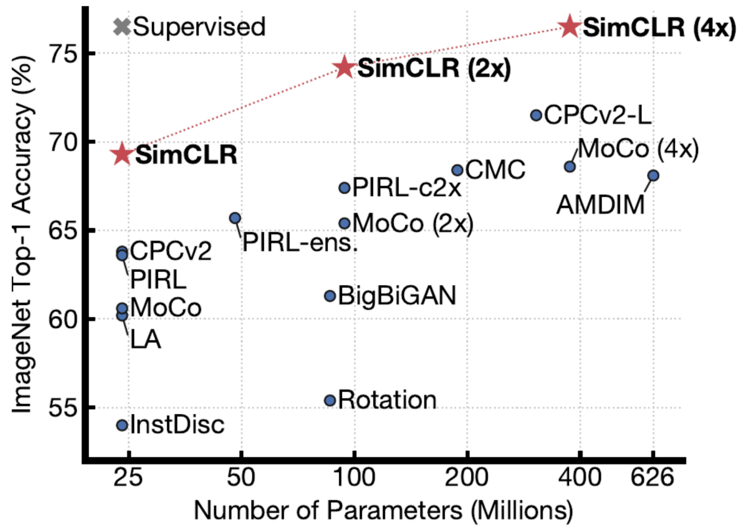
Objective: instance discrimination within batches

Implement with a contrastive loss. For positive pair (i, j):

$$l_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$

positive pair

all pairs



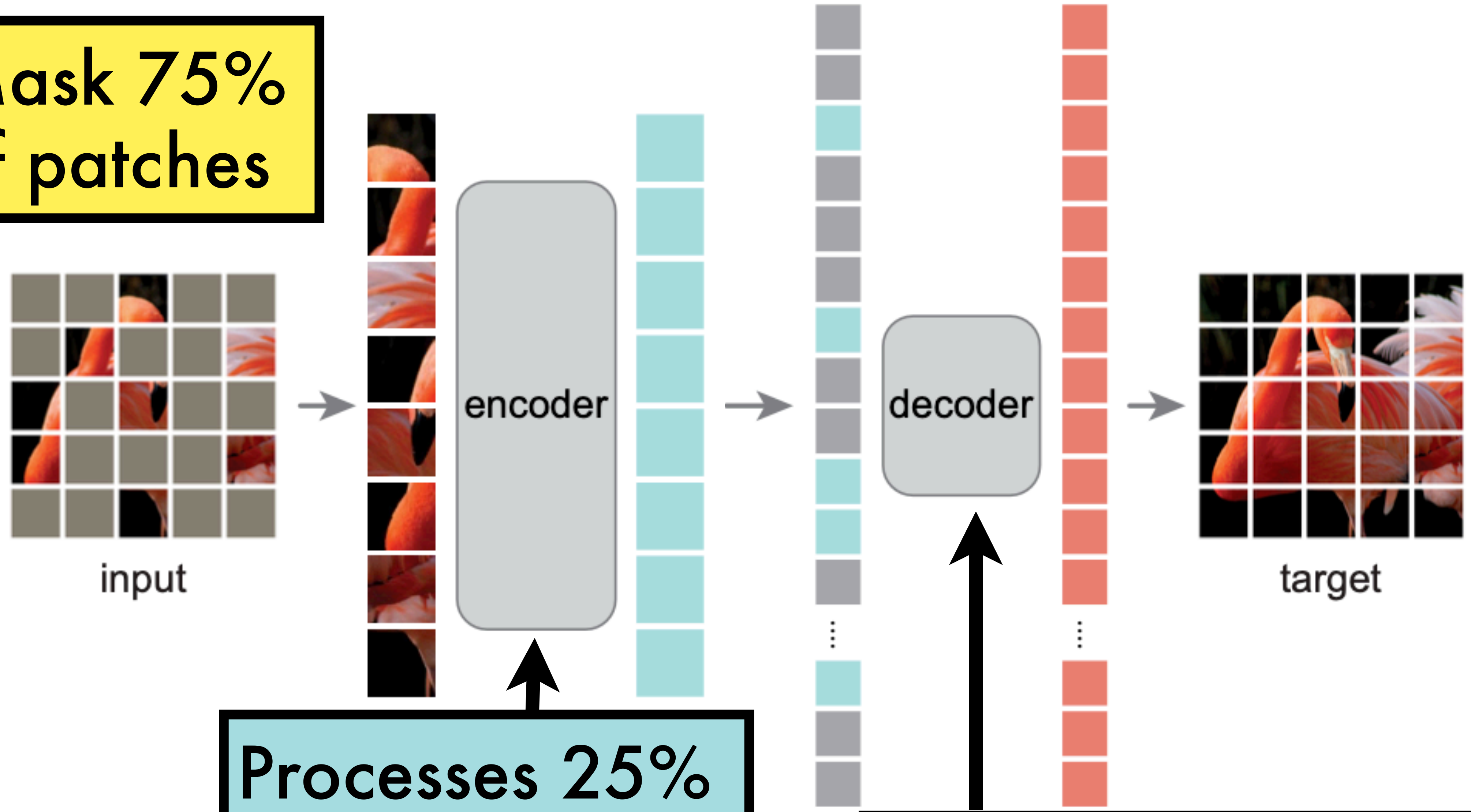
References/Image credits

T. Chen et al., "A simple framework for contrastive learning of visual representations", ICML (2020)

Masked Autoencoders

Masked Autoencoders and scalable learning (He et al., 2022)

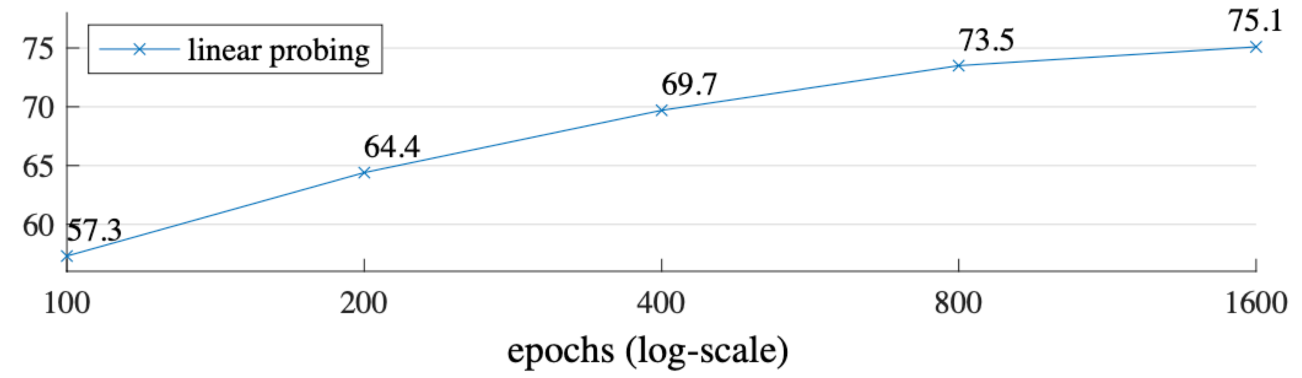
Mask 75% of patches



Processes 25% of patches

Lightweight decoder

Efficient asymmetric encoder-decoder

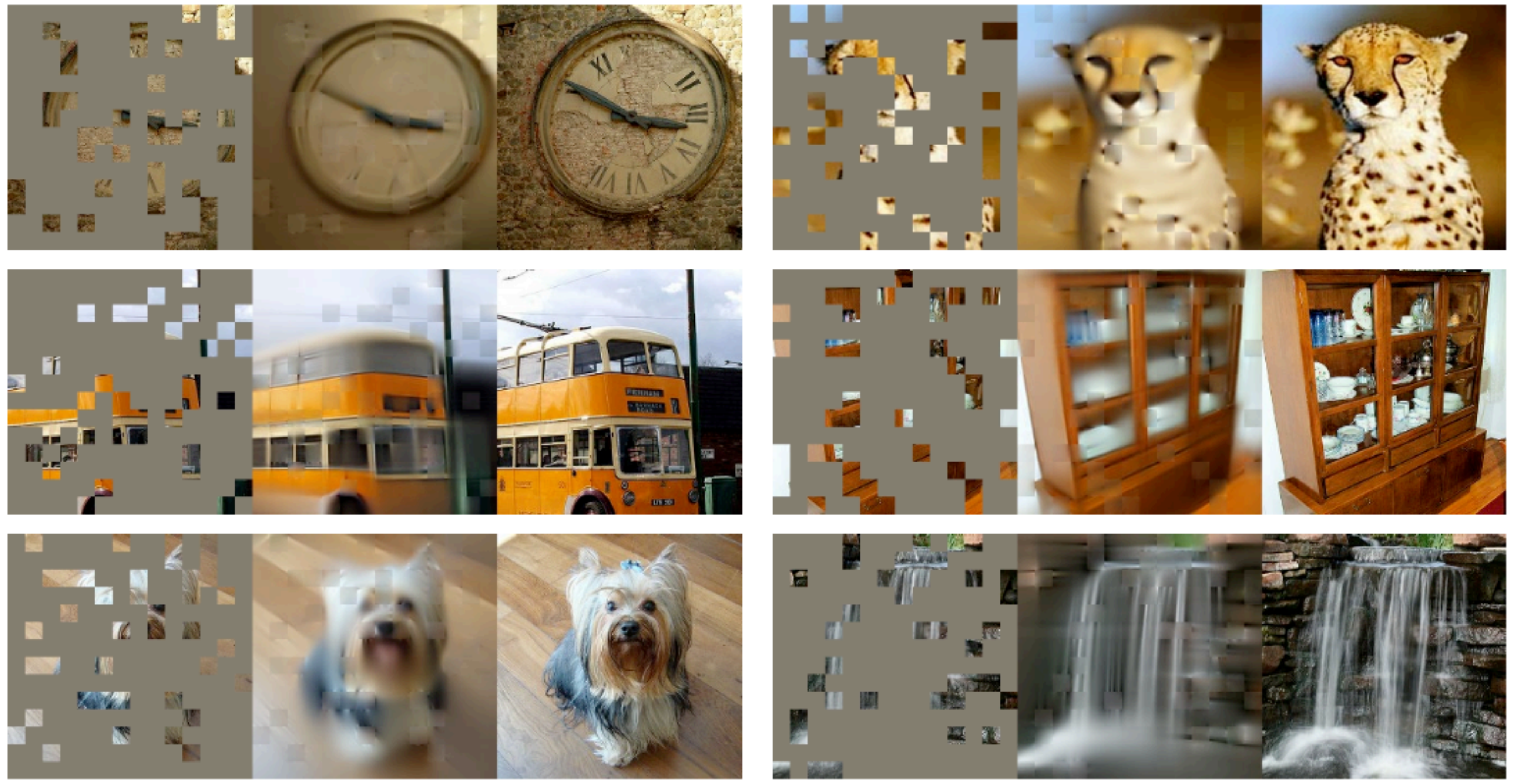


Long training schedules are vital

Idea: Pixel reconstruction with high masking ratios, high-capacity transformers and L2 loss

Reconstructions are blurry, but still drive strong feature learning

input | pred | GT | input | pred | GT



References/Image credits

K. He et al., "Masked autoencoders are scalable vision learners", CVPR (2022)